Chapter 5: Sampling Distributions

Tuesday (3/10/15): 5.1

- Sampling Distributions
- Sample means and the Law of Large Numbers
- The Central Limit Theorem

Thursday (3/12/15): 5.2, Sampling Distributions for Counts and Proportions
Homework: WebAssign due Friday, Excel assignment posted today, due Tuesday after Spring Break.

Parameters and Statistics

Recall:

– A *parameter* is a number that describes an actual characteristic of a population. Assuming we cannot sample the entire population, parameters are never known exactly.

Parameters and Statistics

Recall:

- A *parameter* is a number that describes an actual characteristic of a population. Assuming we cannot sample the entire population, parameters are never known exactly.
- A *statistic* is a number that describes a characteristic of a sample.

We use statistics to estimate parameters.

Parameters and Statistics

Recall:

- A *parameter* is a number that describes an actual characteristic of a population. Assuming we cannot sample the entire population, parameters are never known exactly.
- A *statistic* is a number that describes a characteristic of a sample.

We use statistics to estimate parameters.

Notation for samples vs. populations:

Symbol	Use
μ	*Population* mean
$\bar{x}$	*Sample* mean
σ	*Population* standard deviation
s, s_x	*Sample* standard deviation

Why sampling distributions?

Fundamental idea:

We want to use sample statistics to estimate population parameters. Different random samples will yield different statistics, so we must know the ***distribution*** *of these sample statistics!*

Why sampling distributions?

Fundamental idea:

We want to use sample statistics to estimate population parameters. Different random samples will yield different statistics, so we must know the ***distribution*** *of these sample statistics!*

– Sample statistics will be treated like random variables, and we already know how to find the distribution (PDF) of a random variable (chapter 4).

Why sampling distributions?

Fundamental idea:

We want to use sample statistics to estimate population parameters. Different random samples will yield different statistics, so we must know the ***distribution*** *of these sample statistics!*

– Sample statistics will be treated like random variables, and we already know how to find the distribution (PDF) of a random variable (chapter 4).
– Recall: given a random variable X, its *probability distribution* is a table (for discrete RV's) or function (for continuous RV's) that provides the range of possible outputs of X and their probabilities.

X	2	4	8
$p_X(x)$	1/4	1/2	1/4

Sampling Example

Example: suppose there are 50,000 students at UA. We give everyone a slip of paper with a number written on it. 10,000 students get a 1, 10,000 students get a 2, etc, up to 5.

Sampling Example

Example: suppose there are 50,000 students at UA. We give everyone a slip of paper with a number written on it. 10,000 students get a 1, 10,000 students get a 2, etc, up to 5.

– If we select a student at random and look at their number, this is a random variable X.

Sampling Example

Example: suppose there are 50,000 students at UA. We give everyone a slip of paper with a number written on it. 10,000 students get a 1, 10,000 students get a 2, etc, up to 5.

– If we select a student at random and look at their number, this is a random variable X.

– What is the distribution of X?

X	1	2	3	4	5
$p_X(x)$	1/5	1/5	1/5	1/5	1/5

Sampling Example

Example: suppose there are 50,000 students at UA. We give everyone a slip of paper with a number written on it. 10,000 students get a 1, 10,000 students get a 2, etc, up to 5.

– If we select a student at random and look at their number, this is a random variable X.

– What is the distribution of X?

X	1	2	3	4	5
$p_X(x)$	1/5	1/5	1/5	1/5	1/5

– What is the population mean of X?

$$\mu_X = \sum_{i=1}^{5} p_i x_i = \frac{1}{5}(1+2+3+4+5) = 2.5$$

Sampling Example

Example: suppose there are 50,000 students at UA. We give everyone a slip of paper with a number written on it. 10,000 students get a 1, 10,000 students get a 2, etc, up to 5.

– If we select a student at random and look at their number, this is a random variable X.

– What is the distribution of X?

X	1	2	3	4	5
$p_X(x)$	1/5	1/5	1/5	1/5	1/5

– What is the population mean of X?

$$\mu_X = \sum_{i=1}^{5} p_i x_i = \frac{1}{5}(1+2+3+4+5) = 2.5$$

– What is the population standard deviation?

$$\sigma_X = \sum_{i=1}^{5} p_i (x_i - \mu_X)^2 = \frac{1}{5}\left((1-2.5)^2 + \ldots + (5-2.5)^2\right) = 2.25$$

Sample Means and the Law of Large Numbers

– Now, suppose we select 100 students at random and record their number. You can simulate this in excel using RANDBETWEEN(1,5).

$$1, 5, 2, 2, 5, 4, 3, 1, 3, 5, \ldots$$

Sample Means and the Law of Large Numbers

– Now, suppose we select 100 students at random and record their number. You can simulate this in excel using RANDBETWEEN(1,5).

$$1, 5, 2, 2, 5, 4, 3, 1, 3, 5, \ldots$$

– We think of each of the above numbers as a random variable X_i. In other words, we have 100 random variables!

Sample Means and the Law of Large Numbers

– Now, suppose we select 100 students at random and record their number. You can simulate this in excel using RANDBETWEEN(1,5).

$$1, 5, 2, 2, 5, 4, 3, 1, 3, 5, \ldots$$

– We think of each of the above numbers as a random variable X_i. In other words, we have 100 random variables!
– We can compute the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$, with $n = 100$.

Sample Means and the Law of Large Numbers

– Now, suppose we select 100 students at random and record their number. You can simulate this in excel using RANDBETWEEN(1,5).

$$1, 5, 2, 2, 5, 4, 3, 1, 3, 5, \ldots$$

– We think of each of the above numbers as a random variable X_i. In other words, we have 100 random variables!
– We can compute the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$, with $n = 100$.

Law of large numbers: as the number of samples (n) gets large,
$$\bar{x} \to \mu_X.$$

Sample Means and the Law of Large Numbers

– Now, suppose we select 100 students at random and record their number. You can simulate this in excel using RANDBETWEEN(1,5).

$$1, 5, 2, 2, 5, 4, 3, 1, 3, 5, \ldots$$

– We think of each of the above numbers as a random variable X_i. In other words, we have 100 random variables!
– We can compute the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$, with $n = 100$.

Law of large numbers: as the number of samples (n) gets large,
$$\bar{x} \to \mu_X.$$

– **We are more interested in the variability of $\bar{x}$ across all possible samples of size 100.**

Sample Means and the Law of Large Numbers

– Now, suppose we select 100 students at random and record their number. You can simulate this in excel using RANDBETWEEN(1,5).

$$1, 5, 2, 2, 5, 4, 3, 1, 3, 5, \ldots$$

– We think of each of the above numbers as a random variable X_i. In other words, we have 100 random variables!
– We can compute the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$, with $n = 100$.

Law of large numbers: as the number of samples (n) gets large,
$$\bar{x} \to \mu_X .$$

– **We are more interested in the variability of $\bar{x}$ across all possible samples of size 100.**
– In particular, $\bar{x}$ is a random variable, since it is a linear combination of random variables! **What is its distribution?**

Bias, Variability, Unbiased Estimators

Recall a concept from chapter 3:

If we compute a statistic (such as $\bar{x}$ or s_X) over all possible samples of a certain size, this statistic will vary.

Bias, Variability, Unbiased Estimators

Recall a concept from chapter 3:

If we compute a statistic (such as $\bar{x}$ or s_X) over all possible samples of a certain size, this statistic will vary.

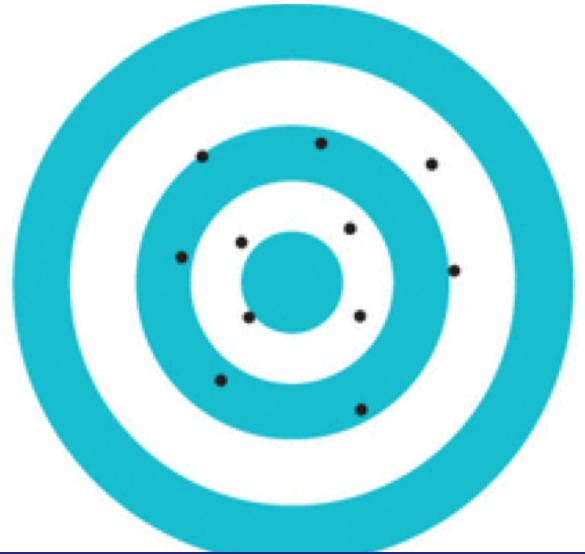
– **Bias** is when the mean of the statistic, over all samples, is *different* from the population parameter. A statistic is an **unbiased estimator** if its mean over all samples is actually *equal* to the population parameter.

Bias, Variability, Unbiased Estimators

Recall a concept from chapter 3:

If we compute a statistic (such as $\bar{x}$ or s_X) over all possible samples of a certain size, this statistic will vary.

- **Bias** is when the mean of the statistic, over all samples, is *different* from the population parameter. A statistic is an **unbiased estimator** if its mean over all samples is actually *equal* to the population parameter.
- **Variability** is how spread out the statistic is over all samples

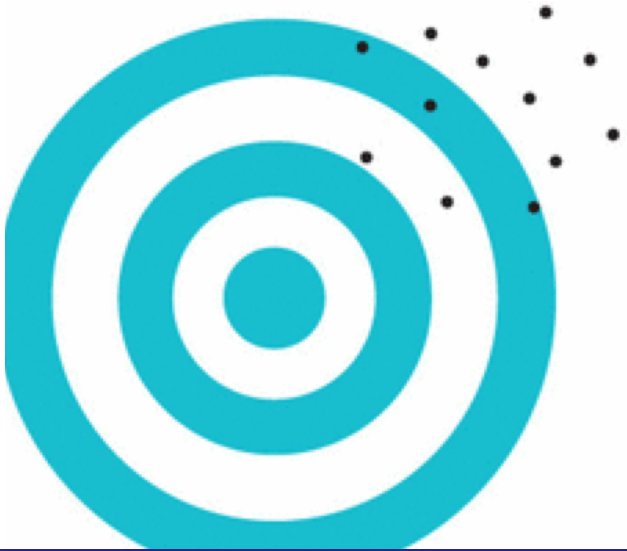


Bias, Variability, Unbiased Estimators

Recall a concept from chapter 3:

If we compute a statistic (such as $\bar{x}$ or s_X) over all possible samples of a certain size, this statistic will vary.

- **Bias** is when the mean of the statistic, over all samples, is *different* from the population parameter. A statistic is an **unbiased estimator** if its mean over all samples is actually *equal* to the population parameter.
- **Variability** is how spread out the statistic is over all samples



Bias, Variability, Unbiased Estimators

Recall a concept from chapter 3:

If we compute a statistic (such as $\bar{x}$ or s_X) over all possible samples of a certain size, this statistic will vary.

- **Bias** is when the mean of the statistic, over all samples, is *different* from the population parameter. A statistic is an **unbiased estimator** if its mean over all samples is actually *equal* to the population parameter.
- **Variability** is how spread out the statistic is over all samples

Mean and standard deviation of the sample mean

Yes, that's right - we are going to compute the *mean of the mean* and the *standard deviation of the mean!* Here is what we mean:

Mean and standard deviation of the sample mean

Yes, that's right - we are going to compute the *mean of the mean* and the *standard deviation of the mean!* Here is what we mean:

– For a given sample size n, **where n is small relative to the population size**, each individual sample is a random variable X_i which is **independent from the others** and **identically distributed**. In other words, the distribution of X_{10} is the same as the distribution of X_{1000}.

Mean and standard deviation of the sample mean

Yes, that's right - we are going to compute the *mean of the mean* and the *standard deviation of the mean!* Here is what we mean:

- For a given sample size n, **where n is small relative to the population size**, each individual sample is a random variable X_i which is **independent from the others** and **identically distributed**. In other words, the distribution of X_{10} is the same as the distribution of X_{1000}.
- In particular, the *population mean and standard deviation* of each X_i is the same:

$$\mu_{X_i} = \mu_X = \mu, \qquad \sigma_{X_i} = \sigma_X = \sigma$$

Another way of saying this: choosing a random sample of 100 *without replacement* from a very large population (e.g. 50,000) will result in the same behavior as if we sampled *with replacement*.

Mean and standard deviation of the sample mean

– Knowing that $\mu_{X_i} = \mu$ for all i, we can compute the mean of $\bar{x}$ (recall that $\bar{x}$ is a random variable!)

$$\mu_{\bar{x}} = E[\bar{x}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\left[\sum_{i=1}^{n} E[X_i]\right] = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}n \cdot \mu = \boxed{\mu}$$

So, this says that $\bar{x}$ is an **unbiased estimator of** μ, since the mean of the *sampling* distribution is the same as the mean of the population!

Mean and standard deviation of the sample mean

– Knowing that $\mu_{X_i} = \mu$ for all i, we can compute the mean of $\bar{x}$ (recall that $\bar{x}$ is a random variable!)

$$\mu_{\bar{x}} = E[\bar{x}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\left[\sum_{i=1}^{n} E[X_i]\right] = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}n \cdot \mu = \boxed{\mu}$$

So, this says that $\bar{x}$ is an **unbiased estimator of** μ, since the mean of the *sampling* distribution is the same as the mean of the population!

– Note: the assumption that n is small compared to the population is crucial. If n is not small compared to population, Law of Large Numbers is more useful.

Mean and standard deviation of the sample mean

– Knowing that $\mu_{X_i} = \mu$ for all i, we can compute the mean of $\bar{x}$ (recall that $\bar{x}$ is a random variable!)

$$\mu_{\bar{x}} = E[\bar{x}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\left[\sum_{i=1}^{n} E[X_i]\right] = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}n\cdot\mu = \boxed{\mu}$$

So, this says that $\bar{x}$ is an **unbiased estimator of** μ, since the mean of the *sampling* distribution is the same as the mean of the population!

– Note: the assumption that n is small compared to the population is crucial. If n is not small compared to population, Law of Large Numbers is more useful.

– We can also compute the sample variance/standard deviation:

$$\sigma^2_{\bar{x}} = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2_{X_i} = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}$$

so, $\boxed{\sigma_{\bar{x}} = \sigma/\sqrt{n}}$

Interpretation

– Recall the setting: we are taking simple random samples of size n from a population of size much bigger than n.

Interpretation

- Recall the setting: we are taking simple random samples of size n from a population of size much bigger than n.
- We can *estimate* the population mean by computing $\bar{x}$. This is an unbiased estimate because if we took $\bar{x}$ for *every possible* sample of size n, the distribution of $\bar{x}$ would be centered around μ.

Interpretation

- Recall the setting: we are taking simple random samples of size n from a population of size much bigger than n.
- We can *estimate* the population mean by computing $\bar{x}$. This is an unbiased estimate because if we took $\bar{x}$ for *every possible* sample of size n, the distribution of $\bar{x}$ would be centered around μ.
- The larger we take n (but not *too* big...), the smaller $\sigma_{\bar{x}}$ gets. In other words, **the distribution of $\bar{x}$ becomes less variable if we take larger samples.**

Interpretation

- Recall the setting: we are taking simple random samples of size n from a population of size much bigger than n.
- We can *estimate* the population mean by computing $\bar{x}$. This is an unbiased estimate because if we took $\bar{x}$ for *every possible* sample of size n, the distribution of $\bar{x}$ would be centered around μ.
- The larger we take n (but not *too* big...), the smaller $\sigma_{\bar{x}}$ gets. In other words, **the distribution of $\bar{x}$ becomes less variable if we take larger samples.**
- Intuition: averaging (computing $\bar{x}$ over a sample *smooths out* the variability, making sample means less variable than individual observations.

Interpretation

- Recall the setting: we are taking simple random samples of size n from a population of size much bigger than n.
- We can *estimate* the population mean by computing $\bar{x}$. This is an unbiased estimate because if we took $\bar{x}$ for *every possible* sample of size n, the distribution of $\bar{x}$ would be centered around μ.
- The larger we take n (but not *too* big...), the smaller $\sigma_{\bar{x}}$ gets. In other words, **the distribution of $\bar{x}$ becomes less variable if we take larger samples.**
- Intuition: averaging (computing $\bar{x}$ over a sample *smooths out* the variability, making sample means less variable than individual observations.
- The question still remains:

What does the distribution of $\bar{x}$ look like?

The distribution of $\bar{x}$

– Using excel, we generate uniform random whole numbers from $\{1, 2, 3, 4, 5\}$ using RANDBETWEEN(1,5). This is like drawing random people from our UA student example - recall that each of 50,000 people had a number between 1 and 5, with the same number of people holding each number.

The distribution of $\bar{x}$

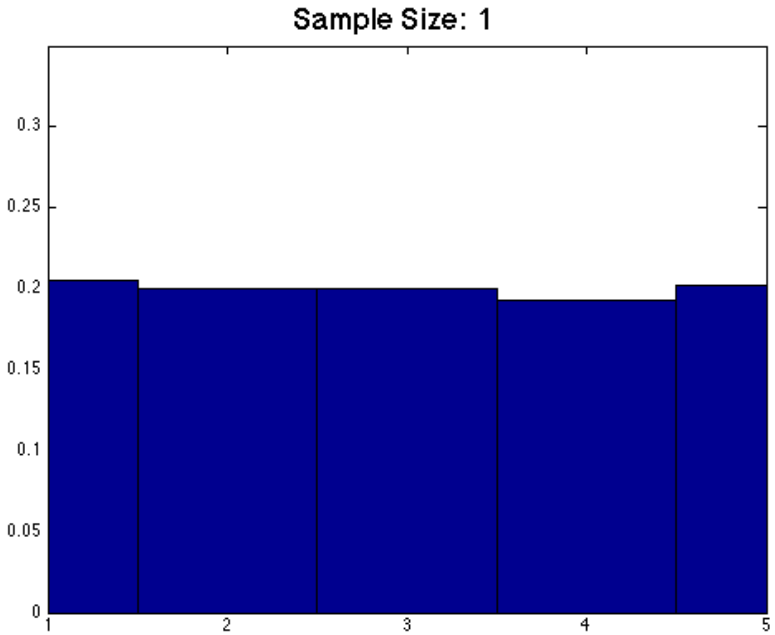
- Using excel, we generate uniform random whole numbers from $\{1, 2, 3, 4, 5\}$ using RANDBETWEEN(1,5). This is like drawing random people from our UA student example - recall that each of 50,000 people had a number between 1 and 5, with the same number of people holding each number.
- To simulate random sampling of size n, we can simply use a range of cells of size n, i.e. A1:A5 or A11:A15 are simple random samples of size 5.

The distribution of $\bar{x}$

- Using excel, we generate uniform random whole numbers from $\{1, 2, 3, 4, 5\}$ using RANDBETWEEN(1,5). This is like drawing random people from our UA student example - recall that each of 50,000 people had a number between 1 and 5, with the same number of people holding each number.
- To simulate random sampling of size n, we can simply use a range of cells of size n, i.e. A1:A5 or A11:A15 are simple random samples of size 5.
- We compute $\bar{x}$ for different random samples of size n, then make a histogram of $\bar{x}$

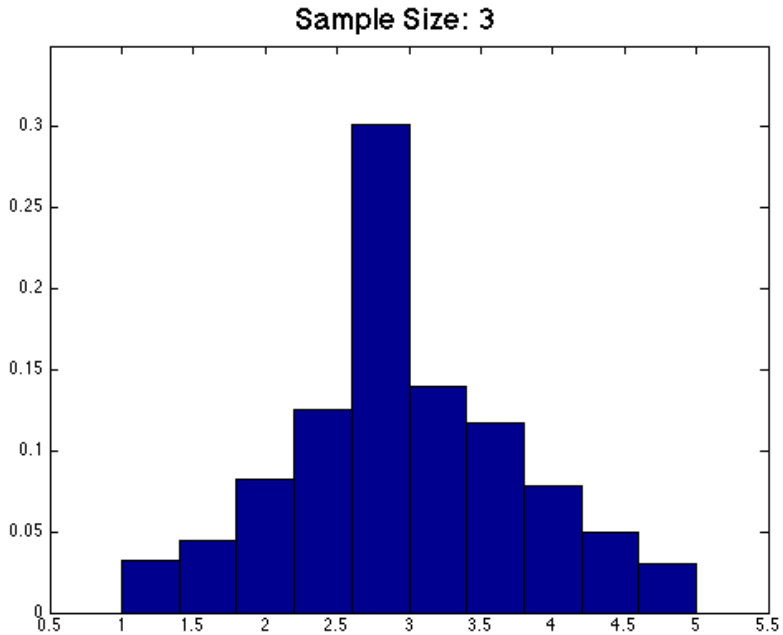
The distribution of $\bar{x}$

We know that for samples of size n, $\bar{x}$ has mean μ (same as the population mean) and standard deviation $\sigma/\sqrt{n}$. The distributions look like:



The distribution of $\bar{x}$

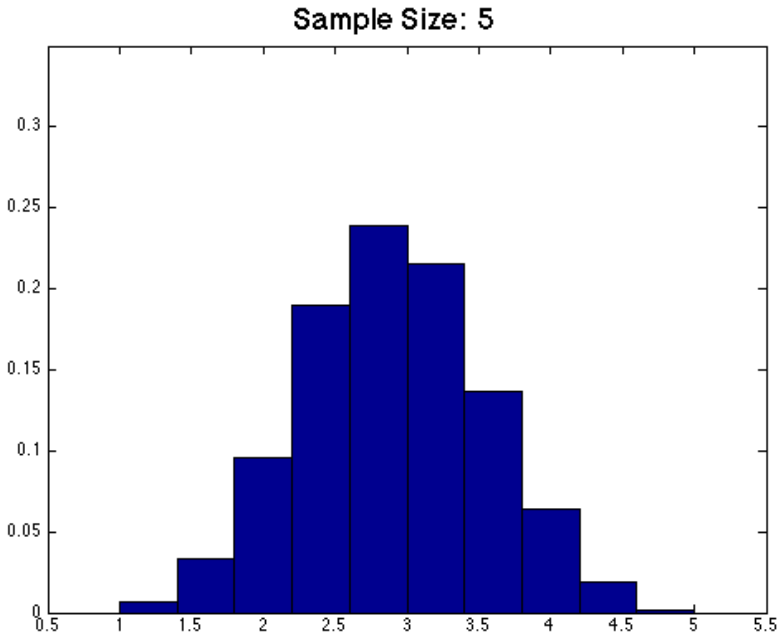
We know that for samples of size n, $\bar{x}$ has mean μ (same as the population mean) and standard deviation $\sigma/\sqrt{n}$. The distributions look like:



These look suspiciously normal!

The distribution of $\bar{x}$

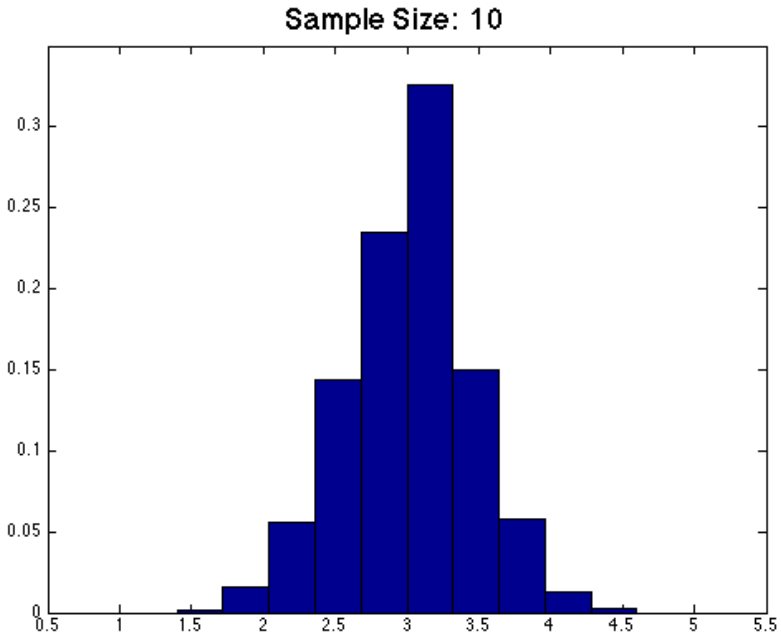
We know that for samples of size n, $\bar{x}$ has mean μ (same as the population mean) and standard deviation $\sigma/\sqrt{n}$. The distributions look like:



These look suspiciously normal!

The distribution of $\bar{x}$

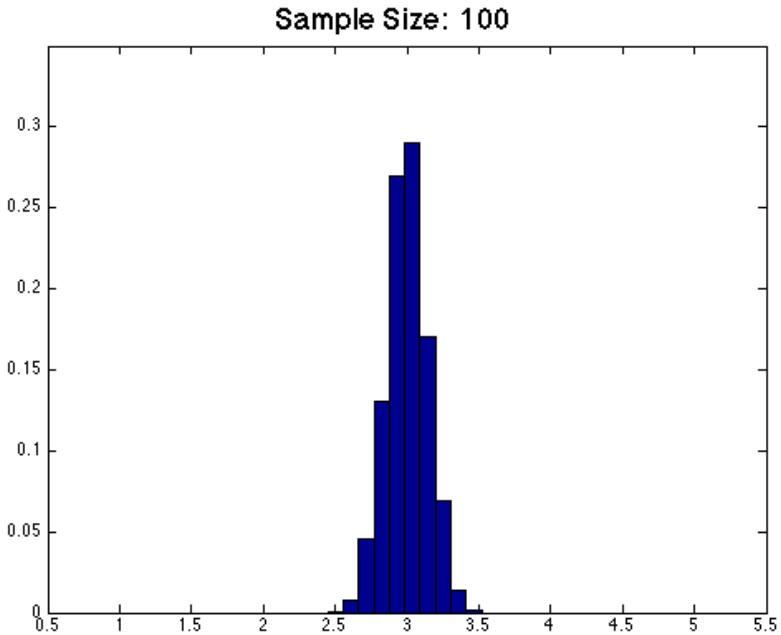
We know that for samples of size n, $\bar{x}$ has mean μ (same as the population mean) and standard deviation $\sigma/\sqrt{n}$. The distributions look like:



These look suspiciously normal!

The distribution of $\bar{x}$

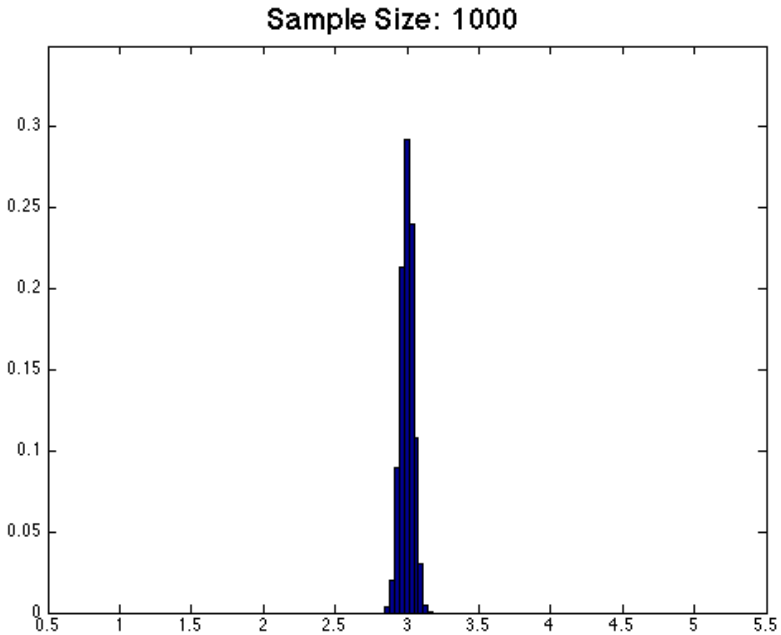
We know that for samples of size n, $\bar{x}$ has mean μ (same as the population mean) and standard deviation $\sigma/\sqrt{n}$. The distributions look like:



These look suspiciously normal!

The distribution of $\bar{x}$

We know that for samples of size n, $\bar{x}$ has mean μ (same as the population mean) and standard deviation $\sigma/\sqrt{n}$. The distributions look like:



These look suspiciously normal!

The Central Limit Theorem

Suppose that $X_1, X_2, \ldots, X_n$ are independent, identically distributed random variables. For instance, drawing random samples from a large population, with $n \ll$ population size, each individual shares the same distribution as X, the population parameter. Then, the distribution of

$$\bar{x} = \frac{1}{n}(X_1 + X_n + \ldots + X_n)$$

*is **approximately normal** (unless the X_i are normal) with mean μ_X and standard deviation $\sigma_X/\sqrt{n}$. If the X_i are normal, then $\bar{x}$ is exactly $N(\mu, \sigma/\sqrt{n})$.*

The Central Limit Theorem

Suppose that $X_1, X_2, \ldots, X_n$ are independent, identically distributed random variables. For instance, drawing random samples from a large population, with $n \ll$ population size, each individual shares the same distribution as X, the population parameter. Then, the distribution of

$$\bar{x} = \frac{1}{n}(X_1 + X_n + \ldots + X_n)$$

is ***approximately normal*** *(unless the X_i are normal) with mean μ_X and standard deviation $\sigma_X/\sqrt{n}$. If the X_i are normal, then $\bar{x}$ is exactly $N(\mu, \sigma/\sqrt{n})$.*

– In other words,

$$\bar{x} \sim N(\mu, \sigma/\sqrt{n}) \qquad \text{(Approximately)}$$

The Central Limit Theorem

Suppose that $X_1, X_2, \ldots, X_n$ are independent, identically distributed random variables. For instance, drawing random samples from a large population, with $n \ll$ population size, each individual shares the same distribution as X, the population parameter. Then, the distribution of

$$\bar{x} = \frac{1}{n}(X_1 + X_n + \ldots + X_n)$$

is ***approximately normal*** *(unless the X_i are normal) with mean μ_X and standard deviation $\sigma_X/\sqrt{n}$. If the X_i are normal, then $\bar{x}$ is exactly $N(\mu, \sigma/\sqrt{n})$.*

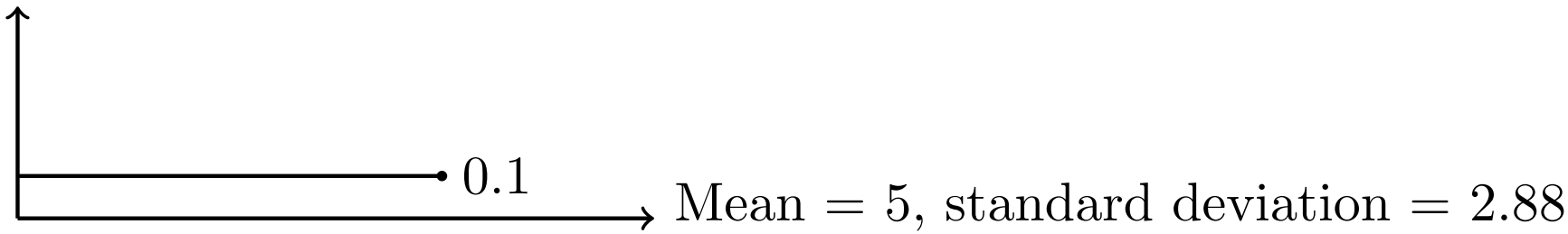
– In other words,

$$\bar{x} \sim N(\mu, \sigma/\sqrt{n}) \qquad \text{(Approximately)}$$

– Note: It does *not matter* what the probability distribution of the variables X_i is! This works with normally distributed RVs, uniformly distributed RVs, and RVs with crazy skewed distributions.

Example

Suppose that we draw random real numbers uniformly from 0 to 10:

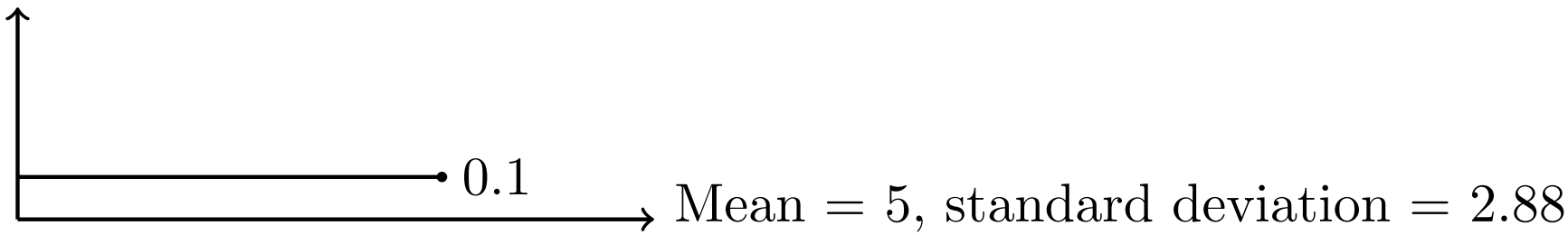


a) What is the probability that a given number drawn is between 1 and 2?

b) Suppose we select 25 numbers randomly. Approximately what is the probability that the *mean* of these numbers is between 1 and 2?

Example

Suppose that we draw random real numbers uniformly from 0 to 10:



a) What is the probability that a given number drawn is between 1 and 2?

b) Suppose we select 25 numbers randomly. Approximately what is the probability that the *mean* of these numbers is between 1 and 2?

Answer:

a) Find the area under the PDF between 1 and 2.

b) Using the CLT, we know that $\bar{x} \sim N(5, 2.88/\sqrt{25})$, so we use

normalcdf(1,2,5,2.88/5)

Example

Suppose ACT scores across all incoming freshmen in the US are normally distributed with a mean of 20.8 and a standard deviation of 4.8.

a) Are the given means and standard deviations for the population or for a sample?

b) What is the probability that a randomly selected student had a score higher than 30?

c) What is the probability that 5 randomly selected students have their *average* score higher than 30?

Example

Suppose ACT scores across all incoming freshmen in the US are normally distributed with a mean of 20.8 and a standard deviation of 4.8.

a) Are the given means and standard deviations for the population or for a sample?

b) What is the probability that a randomly selected student had a score higher than 30?

c) What is the probability that 5 randomly selected students have their *average* score higher than 30?

Answers:

a) These are for the *population*

b) Since this random variable is normally distributed, we can use normalcdf(30,1000,20.8,4.8) = $\boxed{0.028}$ (2 sig figs). As usual, the 1000 is just a 'large' number to indicate that we want all numbers to the right of 30.

Example, cont'd

Solution to part c): by the central limit theorem, we know that

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{5}}\right)$$

Since our original random variable is normal, this is in fact exact - $\bar{x}$ will have *exactly* a normal distribution with mean $\mu = 20.8$ and standard deviation $\sigma/\sqrt{5}$ (5 is the sample size). Thus the probability that 5 randomly selected students has an average score higher than 30 is found with normalcdf(30,1000,20.8,4.8/$\sqrt{5}$) = $\boxed{9.11 \cdot 10^{-6}}$

Review of LLN and CLT

Assumptions:

1) I have n **independent identically distributed** random variables $X_1, \ldots, X_n$, each with mean μ and standard deviation σ.

In most examples, the situation is this: each X_i represents a variable for an individual drawn from a very large population, so that the collection $\{X_1, \ldots, X_n\}$ models a sample of size n. For example, each X_i might be a randomly selected person's age or height or blood O2 level.

Review of LLN and CLT

Assumptions:

1) I have n **independent identically distributed** random variables $X_1, \ldots, X_n$, each with mean μ and standard deviation σ.

   In most examples, the situation is this: each X_i represents a variable for an individual drawn from a very large population, so that the collection $\{X_1, \ldots, X_n\}$ models a sample of size n. For example, each X_i might be a randomly selected person's age or height or blood O2 level.

2) The random variables must have **finite mean and standard deviation** (not an issue in this class, but it does come up!)

Review of LLN and CLT

Assumptions:

1) I have n **independent identically distributed** random variables $X_1, \ldots, X_n$, each with mean μ and standard deviation σ.

   In most examples, the situation is this: each X_i represents a variable for an individual drawn from a very large population, so that the collection $\{X_1, \ldots, X_n\}$ models a sample of size n. For example, each X_i might be a randomly selected person's age or height or blood O2 level.

2) The random variables must have **finite mean and standard deviation** (not an issue in this class, but it does come up!)

Define the *sample mean* as the random variable

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Review of LLN and CLT

Conclusions:

1) Law of large numbers: $\bar{X} \to \mu$. Loosely speaking we mean that as we collect larger and larger samples, the sample mean 'goes to' the population mean.
2) Central limit theorem: *the variation of* $\bar{X}$ *around* μ *follows a normal distribution.* More precisely,

$$\bar{X} \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

– This means that if we collected many samples of a fixed size n, we would expect the sample means of these samples to be clustered around the actual population mean. The variability decreases with larger samples, **as long as we still satisfy the IID assumption**.

5.2: Repeated Trials and the Binomial Distribution

- A *Bernoulli trial* is an experiment where the outcome is one of two possibilities - i.e. 'yes' or 'no', 'heads' or 'tails', etc.
- When we repeat Bernoulli trials and record the number of positive results, we are in a *Binomial* setting.

5.2: Repeated Trials and the Binomial Distribution

- A *Bernoulli trial* is an experiment where the outcome is one of two possibilities - i.e. 'yes' or 'no', 'heads' or 'tails', etc.
- When we repeat Bernoulli trials and record the number of positive results, we are in a *Binomial* setting.

More precisely, we need:

- A binary experiment (outcome is one of two options)
- Each 'trial' is independent
- We fix a number of trials at the start
- Each trial has the same probability p of success and probability $1 - p$ of failure.

5.2: Repeated Trials and the Binomial Distribution

- A *Bernoulli trial* is an experiment where the outcome is one of two possibilities - i.e. 'yes' or 'no', 'heads' or 'tails', etc.
- When we repeat Bernoulli trials and record the number of positive results, we are in a *Binomial* setting.

More precisely, we need:

- A binary experiment (outcome is one of two options)
- Each 'trial' is independent
- We fix a number of trials at the start
- Each trial has the same probability p of success and probability $1 - p$ of failure.

Example: flip a fair coin 5 times and count the number of heads. What is p?

5.2: Repeated Trials and the Binomial Distribution

- A *Bernoulli trial* is an experiment where the outcome is one of two possibilities - i.e. 'yes' or 'no', 'heads' or 'tails', etc.
- When we repeat Bernoulli trials and record the number of positive results, we are in a *Binomial* setting.

More precisely, we need:

- A binary experiment (outcome is one of two options)
- Each 'trial' is independent
- We fix a number of trials at the start
- Each trial has the same probability p of success and probability $1 - p$ of failure.

Example: flip a fair coin 5 times and count the number of heads. What is p? *Answer:* $p = 1/2$ *since the coin is fair.*

Example

Suppose we roll a (fair) 6-sided die. A 'success' (S) will be if the die shows a 6. Any other number is considered a fail (F) We will roll the die 3 times and count the number of successes.

Example

Suppose we roll a (fair) 6-sided die. A 'success' (S) will be if the die shows a 6. Any other number is considered a fail (F) We will roll the die 3 times and count the number of successes.

– What is the probability of a success on a given trial?

Example

Suppose we roll a (fair) 6-sided die. A 'success' (S) will be if the die shows a 6. Any other number is considered a fail (F) We will roll the die 3 times and count the number of successes.

– What is the probability of a success on a given trial? *Answer:* $p = 1/6$.

Example

Suppose we roll a (fair) 6-sided die. A 'success' (S) will be if the die shows a 6. Any other number is considered a fail (F) We will roll the die 3 times and count the number of successes.

- What is the probability of a success on a given trial? *Answer:* $p = 1/6$.
- What is the probability of the following: SFF?

Example

Suppose we roll a (fair) 6-sided die. A 'success' (S) will be if the die shows a 6. Any other number is considered a fail (F) We will roll the die 3 times and count the number of successes.

- What is the probability of a success on a given trial? *Answer:* $p = 1/6$.
- What is the probability of the following: SFF? *Answer:* $Pr(SFF) = (1/6) \cdot (5/6) \cdot (5/6) = \frac{25}{216} \approx 0.116$

Example

Suppose we roll a (fair) 6-sided die. A 'success' (S) will be if the die shows a 6. Any other number is considered a fail (F) We will roll the die 3 times and count the number of successes.

- What is the probability of a success on a given trial? *Answer:* $p = 1/6$.
- What is the probability of the following: SFF? *Answer:* $Pr(SFF) = (1/6) \cdot (5/6) \cdot (5/6) = \frac{25}{216} \approx 0.116$
- What is the probability of getting exactly one success in 3 rolls?

Example

Suppose we roll a (fair) 6-sided die. A 'success' (S) will be if the die shows a 6. Any other number is considered a fail (F) We will roll the die 3 times and count the number of successes.

- What is the probability of a success on a given trial? *Answer:* $p = 1/6$.
- What is the probability of the following: SFF? *Answer:* $Pr(SFF) = (1/6) \cdot (5/6) \cdot (5/6) = \frac{25}{216} \approx 0.116$
- What is the probability of getting exactly one success in 3 rolls? *Answer: there are 3 different ways to get one success:* SFF, FSF, FFS. *Each has the same probability,* $25/216$. *So, the probability of getting exactly one success is*

$$3\frac{25}{216} = \frac{75}{216} \approx 0.347$$

Example

Suppose flip an unfair coin 100 times. The probability of heads is 0.4. What is the probability that we get exactly 20 heads in 100 flips?

Example

Suppose flip an unfair coin 100 times. The probability of heads is 0.4. What is the probability that we get exactly 20 heads in 100 flips?
Solution: Let X be the number of heads in 100 flips. The probability of getting exactly 20 heads will be

$$P(X = 20) = N 0.4^{20} 0.6^{80}$$

where N is the number of *different ways* to get 20 heads. We can find N using the 'choose' function:

$$\binom{100}{20} \qquad \text{('100 choose 20')}$$

The Binomial Distribution

When we have repeated Bernoulli trials that satisfy the conditions for a binomial setting (independence, fixed number of trials n, fixed probability of success p), the distribution for the number of success is the *Binomial Distribution*:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

The Binomial Distribution

When we have repeated Bernoulli trials that satisfy the conditions for a binomial setting (independence, fixed number of trials n, fixed probability of success p), the distribution for the number of success is the *Binomial Distribution*:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Intuitively,

- p^k is the probability of getting k successes
- $(1-p)^{n-k}$ is the probability of the rest $(n-k)$ being failures
- $p^k(1-p)^{n-k}$ is the probability of getting exactly k success **and** $n-k$ failures
- $\binom{n}{k}$ is the *number of different ways this can happen.*

The Binomial Distribution

When we have repeated Bernoulli trials that satisfy the conditions for a binomial setting (independence, fixed number of trials n, fixed probability of success p), the distribution for the number of success is the *Binomial Distribution*:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Intuitively,

- p^k is the probability of getting k successes
- $(1-p)^{n-k}$ is the probability of the rest $(n-k)$ being failures
- $p^k(1-p)^{n-k}$ is the probability of getting exactly k success **and** $n-k$ failures
- $\binom{n}{k}$ is the *number of different ways this can happen.*

The number $\binom{n}{k}$ is the binomial coefficient:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \quad n! = n \cdot (n-1) \cdots 2 \cdot 1$$

Computing Binomial Coefficients

– On the TI-84, you can compute $\binom{n}{k}$ by inputting n, then selecting MATH:PRB:nCr, then inputting k. It will look like this:

5 nCr 3

Computing Binomial Coefficients

– On the TI-84, you can compute $\binom{n}{k}$ by inputting n, then selecting MATH:PRB:nCr, then inputting k. It will look like this:

5 nCr 3

– You can compute binomial coefficients by hand using Pascal's triangle:

$$\begin{array}{ccccccccccc} & & & & & 1 & & & & & \\ & & & & 1 & & 1 & & & & \\ & & & 1 & & 2 & & 1 & & & \\ & & 1 & & 3 & & 3 & & 1 & & \\ & 1 & & 4 & & 6 & & 4 & & 1 & \\ 1 & & 5 & & 10 & & 10 & & 5 & & 1 \end{array}$$

Example of Binomial Distribution?

– Example: suppose we are going to test a lot of 1000 climbing ropes for strength. We select 5 ropes at random and test them. Suppose the probability of a rope passing a test is 0.999 (i.e. 999/1000 ropes will pass). Define the random variable X to be the number of ropes that **fail** the test in our sample.

Example of Binomial Distribution?

- Example: suppose we are going to test a lot of 1000 climbing ropes for strength. We select 5 ropes at random and test them. Suppose the probability of a rope passing a test is 0.999 (i.e. 999/1000 ropes will pass). Define the random variable X to be the number of ropes that **fail** the test in our sample.
- Is this an example of a binomial setting?

Example of Binomial Distribution?

- Example: suppose we are going to test a lot of 1000 climbing ropes for strength. We select 5 ropes at random and test them. Suppose the probability of a rope passing a test is 0.999 (i.e. 999/1000 ropes will pass). Define the random variable X to be the number of ropes that **fail** the test in our sample.
- Is this an example of a binomial setting?
- *Answer: No, not quite - each trial is not independent since the population gets smaller each time.*

Example of Binomial Distribution?

- Example: suppose we are going to test a lot of 1000 climbing ropes for strength. We select 5 ropes at random and test them. Suppose the probability of a rope passing a test is 0.999 (i.e. 999/1000 ropes will pass). Define the random variable X to be the number of ropes that **fail** the test in our sample.
- Is this an example of a binomial setting?
- *Answer: No, not quite - each trial is not independent since the population gets smaller each time.*
- What is $P(X = 0)$?

$$P(X = 0) = \frac{999}{1000} \cdot \frac{998}{999} \cdot \frac{997}{998} \cdot \frac{996}{997} \cdot \frac{995}{996} = 0.995$$

Example of Binomial Distribution?

- Example: suppose we are going to test a lot of 1000 climbing ropes for strength. We select 5 ropes at random and test them. Suppose the probability of a rope passing a test is 0.999 (i.e. 999/1000 ropes will pass). Define the random variable X to be the number of ropes that **fail** the test in our sample.
- Is this an example of a binomial setting?
- *Answer: No, not quite - each trial is not independent since the population gets smaller each time.*
- What is $P(X = 0)$?

$$P(X = 0) = \frac{999}{1000} \cdot \frac{998}{999} \cdot \frac{997}{998} \cdot \frac{996}{997} \cdot \frac{995}{996} = 0.995$$

- Try computing this again using a binomial distribution.

$$P(X = 0) \approx \binom{5}{0} 0.001^0 (0.999)^5 \approx 0.99500999$$

Sampling Distribution for a 'Count'

Suppose we sample from a population and count how many individuals satisfy a condition. For example, polling for a yes/no vote on an initiative. We are interested in the distribution of the number of 'successes'.

If we select a simple random sample of size n from a population of size much larger than n, where the probability of success is known to be p, the number of successes k in the sample is ***approximately*** *binomial with parameters p and n:*

$$P(X = k) \approx \binom{n}{k} p^k (1-p)^{n-k}$$

Mean and Standard Deviation of a Binomial RV

The mean and standard deviation of a binomial RV are:

$$\mu = np$$
$$\sigma = \sqrt{np(1-p)}$$

Derivation is tedious, uses the *binomial theorem*:

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

Approximating Binomial Distributions with Normal

If the number of trials is large compared to either the probability of success or failure (say both np and $n(1-p)$ are bigger than 10), but still small compared to the population size, we have

$$\text{Binom}(n,p) \approx N(np, \sqrt{np(1-p)})$$

[1] see Gallup.com

Approximating Binomial Distributions with Normal

If the number of trials is large compared to either the probability of success or failure (say both np and $n(1-p)$ are bigger than 10), but still small compared to the population size, we have

$$\text{Binom}(n,p) \approx N(np, \sqrt{np(1-p)})$$

Example: A survey asks a nationwide random sample of 2500 adults their opinion on same-sex marriage. *Suppose* that 55% of Americans [1] actually support same-sex marriage. Estimate the probability that 1400 people **or more** out of the 2500 support same-sex marriage.

[1] see Gallup.com

Approximating Binomial Distributions with Normal

If the number of trials is large compared to either the probability of success or failure (say both np and $n(1-p)$ are bigger than 10), but still small compared to the population size, we have

$$\text{Binom}(n,p) \approx N(np, \sqrt{np(1-p)})$$

Example: A survey asks a nationwide random sample of 2500 adults their opinion on same-sex marriage. *Suppose* that 55% of Americans [1] actually support same-sex marriage. Estimate the probability that 1400 people **or more** out of the 2500 support same-sex marriage.
– Check for validity of normal approximation:
$np = 2500 \cdot 0.55 = 1375 \gg 10$, $n(1-p) = 1125 \gg 10$.

[1] see Gallup.com

Approximating Binomial Distributions with Normal

If the number of trials is large compared to either the probability of success or failure (say both np and $n(1-p)$ are bigger than 10), but still small compared to the population size, we have

$$\text{Binom}(n,p) \approx N(np, \sqrt{np(1-p)})$$

Example: A survey asks a nationwide random sample of 2500 adults their opinion on same-sex marriage. *Suppose* that 55% of Americans [1] actually support same-sex marriage. Estimate the probability that 1400 people **or more** out of the 2500 support same-sex marriage.

– Check for validity of normal approximation:

$np = 2500 \cdot 0.55 = 1375 \gg 10$, $n(1-p) = 1125 \gg 10$.

– Compute $P(X \geq 1400)$ using z-scores or normalcdf:

$$\mu = np = 1375, \qquad \sigma = \sqrt{np(1-p)} \approx 24.875$$

$$P(X \geq 1400) \approx \text{normalcdf}(1400, 10000, 1375, 24.875) \approx 0.157$$

[1] see Gallup.com

Sample Proportions

Sometimes, we care more about the *proportion* of a population that satisfies some condition rather than the *number* of people. For example 'what percentage of people drink coffee every day'?

$$\hat{p} = \frac{\text{number of successes in sample}}{\text{size of sample}} = \frac{X}{n}$$

Sample Proportions

Sometimes, we care more about the *proportion* of a population that satisfies some condition rather than the *number* of people. For example 'what percentage of people drink coffee every day'?

$$\hat{p} = \frac{\text{number of successes in sample}}{\text{size of sample}} = \frac{X}{n}$$

– $0 \leq \hat{p} \leq 1$, whereas $0 \leq X \leq n$.

Sample Proportions

Sometimes, we care more about the *proportion* of a population that satisfies some condition rather than the *number* of people. For example 'what percentage of people drink coffee every day'?

$$\hat{p} = \frac{\text{number of successes in sample}}{\text{size of sample}} = \frac{X}{n}$$

- $0 \leq \hat{p} \leq 1$, whereas $0 \leq X \leq n$.
- If X is binomial (or is approximately binomial), then

$$\mu_{\hat{p}} = p$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Sample Proportions

Sometimes, we care more about the *proportion* of a population that satisfies some condition rather than the *number* of people. For example 'what percentage of people drink coffee every day'?

$$\hat{p} = \frac{\text{number of successes in sample}}{\text{size of sample}} = \frac{X}{n}$$

- $0 \leq \hat{p} \leq 1$, whereas $0 \leq X \leq n$.
- If X is binomial (or is approximately binomial), then

$$\mu_{\hat{p}} = p$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

- Careful to distinguish $\hat{p}$, which is a sample proportion, from p, which is a probability. $\hat{p}$ is like an approximation of p.

Normal Approximation for Proportion

Making the binomial assumption again - i.e. for sample sizes much smaller than the population size (book uses population = 20*sample as a rule-of-thumb), we can use a normal distribution to approximate a proportion:

$$\hat{p} \approx N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

Normal Approximation for Proportion

Making the binomial assumption again - i.e. for sample sizes much smaller than the population size (book uses population = 20*sample as a rule-of-thumb), we can use a normal distribution to approximate a proportion:

$$\hat{p} \approx N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

Example: suppose we toss a fair coin 500 times. Let X be the number of heads.

a) Find the mean and standard deviation of X

b) Find the mean and standard deviation of $\hat{p}$, the *proportion* of heads out of 500 tosses.

c) Find the probability that between 45% and 50% of the coins showed heads.